Gujarati Text Recognition: A Review

Khushali B. Kathiriya
Department of Information Technology
Dharamsinh Desai University
Nadiad, India
khushipatel2412@gmail.com

Mukesh M. Goswami
Department of Information Technology
Dharamsinh Desai University
Nadiad, India
mukesh.goswami@gmail.com

***Abstract*— Various commercial OCR systems are available for the western scripts. But there is no sufficient work for Indian scripts including Gujarati script. On the other hand, there are few OCR available for different Indian scripts except for Gujarati script. This paper presents a survey of text recognition techniques for Gujarati script. This survey is classified broadly based on Gujarati script. This paper is the result of efforts in two directions namely printed Gujarati documents, and handwritten Gujarati documents.**

Keywords*— *Text Recognition, Online Recognition, Offline Recognition, Optical Character Recognition (OCR), Word spotting(Word matching), Gujarati script.

I. INTRODUCTION

India is a multilingual and multiscript country. There are 22 official languages in India, which is written in 12 different scripts. A large portion of the Indian scripts is derived from the Brahmi script namely Kashmiri, Gurumukhi, Devanagari, Gujarati, Bengali, Oriya, Kannada, Tamil, Telugu, Malayalam. Most of the scripts are written in left to right. [1]

Gujarati is an Indo-Aryan family, which is a subpart of the Indo-European languages. Gujarati is the official language of Gujarat. There are about 65.5 million speakers of Gujarati Worldwide (CIA,2011 [2]). The Gujarati language has rich cultural heritage and literature including printed as well as handwritten documents of Mahatma Gandhi, Poems from Narasimha Mehta, Shikshapatri and letters of god Swaminarayan.

Huge numbers of printed and handwritten documents are available in Gujarati script. It necessary to preserve such documents in the digital format from a historical and legal perspective as well as efficient dissemination. Scanning is one of the best approach to convert documents into digital form. However, editing, searching and retrieving information in these scanned document images is difficult. Therefore, from the scanned document, retrieving the information is an important task. There are mainly two approaches for the IR (Information Retrieval) from document namely Recognition based and Recognition free approach.

The Recognition based approach uses OCR (Optical Character Recognition) system to convert document image into a text (ASCII) document. Recognition free approach considers a word image as a query image and performs the IR task by directly comparing the query word image with the document word images.

A. Overview of the Text Recognition system

Text analysis problem is considered as, 1) Text Recognition, and 2) Text Matching. In text recognition task, from handwritten and printed documents recognizing the word/character is possible in two ways: 1) Off-line word recognition, 2) Online word recognition (Shown in fig.1).

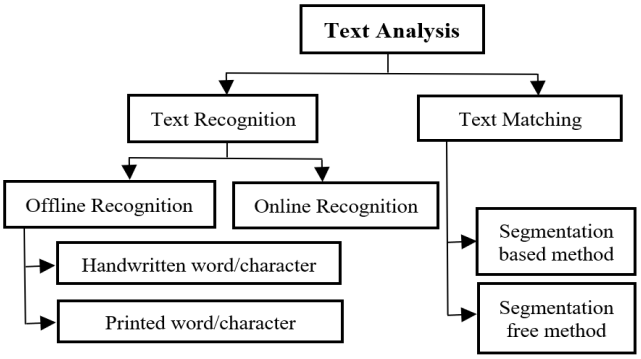

Fig. 1. Types of word/character recognition

Offline text recognition deals with the recognition of words after it has written by people, generally on a paper or sheet. Offline text recognition refers to the process of recognizing the text that is scanned from the paper (or sheet) and stored digitally in formats such as .pdf, .jpeg, .png, .bmp, etc. In online text recognition, the write up is done using digital pen on electronic notepad/ tablet/laptop(touch) [3] [4]. Where in text matching retrieving information without recognizing characters/words explicitly [5].

B. The architecture of Optical Character Recognition

OCR technique is used to convert the scanned documents into a machine-editable text format. The input of the OCR system is in image format and the output in machine editable, searchable and translation format.

There are various steps involved in text recognition of an image (Shown in fig.2), 1) Pre-processing is a step to remove the noise and correct the format of the documents. It is applied to characters before extracting the features and performing classification on them. It has many operations namely skew correction, normalization, noise removal, binarization, thinning, thickening, etc. 2) Segmentation is performed to segment a character in the document. First documents are segmented to the line, lines are segmented to words, and words are further segmented to the character. 3) Feature extraction is used which makes classification of patterns easy just by using the formal procedure. In numerals as well as characters there are certain types of parameters which can be extracted by using different feature extraction techniques. 4) Classification is the main decision-making stage of an Optical Recognition System and uses the feature extracted in the previous stage to identify the text segment according to the present rule. According to extracted features decision can be made through some kind of decision rules.

978-1-5386-8190-9/19/$31.00 © 2019 IEEE

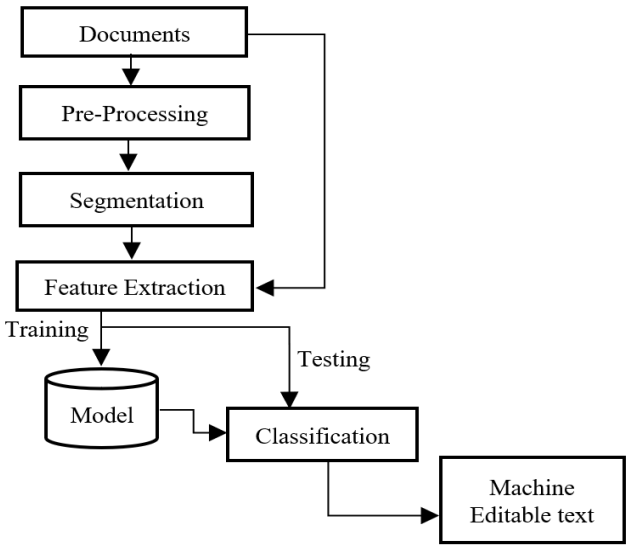

Fig. 2. Components of the OCR system

The rest of the paper is organized as follows: Section II provides a brief introduction about the Gujarati script and properties of it. Section III describes an overview of the Gujarati script in two subsections as Recognition of printed documents and Recognition of handwritten documents. Section IV describes the analysis of literature review and database information. The conclusion is illustrated in Section V.

II. Properties of Gujarati Script

The Gujarati language is written in Gujarati script, which is written from left to right. The Gujarati language's character set consist of 34 constants, 10 numbers, 12 vowels (shown in fig.4). Gujarati language also has conjuncts and join characters (Shown in fig.3) [6].

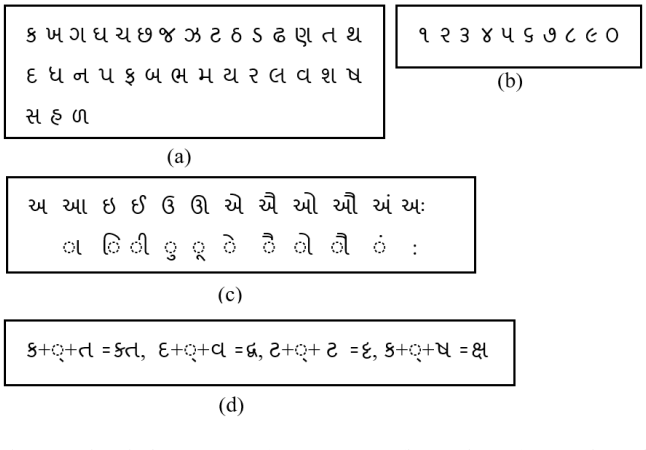

Fig. 3. Gujarati character set: (a)Consonants, (b)Numbers, (c)Vowels, and (d)Conjuncts

Gujarati text is divided into three parallel lines (and zones): 1) upper zone, 2) middle zone, and 3) lower zone. The upper and lower both zones contain modifier symbols, Whereas the middle zone contains base characters and conjuncts (Shown fig. 4) [6].

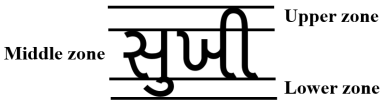

Fig. 4. Upper, middle, and lower zones in Gujarati word

III. A Literature Review of Gujarati Script

Various OCR systems are available for the Indian scripts. However, the majority of work done for recognition of some Indian scripts has described in [3]. But very little work is seen in the literature on recognition of Gujarati scripts.

A. Recognition of printed documents

The first reference found in 1999 by A. Antani and L. Agnihotri [27] described the classification of a subset of printed Gujarati characters. Many researchers have used first and higher order as moment base features with K-nearest neighbor and Minimum Euclidean distance classifiers. The accuracy obtained was 67% on a small dataset. They have collected printed Gujarati character from various sites on the Internet. The training dataset consists of 10 classes, and each class having 10 samples. Where in tested dataset consists of 30 samples of each character.

In 2007, J. Dholakia et al. [7] described classification on printed Gujarati text recognition system. They have used wavelet features with K-nearest neighbor classifier and general regression neural network (GRNN). The overall accuracy claimed was 96-97% on a database of 4,173 symbols and 119 classes.

In 2011, M. M. Goswami et al. [8] have used a Self-Organizing Map (SOM) projection with KNN Classifier. Researchers have generated the small dataset for the Gujarati characters around 3000 symbols in 32 middle-zone character classes. The overall accuracy claimed was 84% on generated characters dataset.

In 2012, M. Chaudhary et al. [9] made a system to recognize similar looking printed Gujarati characters. They used MLP with BPNN as classifiers and the accuracy claimed was 96% on the generated dataset. Their proposed method contained data scanned copies from different newspaper having different font styles and sizes.

In 2013, P. Solanki and M. Bhatt [10] described the work on printed Gujarati documents. Their proposed method has feature extraction using PCA and classifiers as Hopfield Neural Network. The accuracy obtained was 93% on a small dataset.

In 2014, E. Hassan et al. [11] used MKL (Multiple Kernel Learning) based on SVM classifier with Shape Descriptor (SD), HOG (Histogram of Oriented Gradients), Feature Map (FM), Modifiers Shape Descriptor (MSD) features. The overall accuracy was 97-98% on a database of 16000 symbols from 250 classes of Gujarati printed characters.

In 2017 M. M. Goswami and S. K. Mitra [12] described the work on printed Gujarati Character using High-level strokes features. The proposed method was tested on printed Gujarati character dataset consisting of 12000 samples from 42 different classes. They used classifiers as K-nearest neighbor with shape similarity. The overall accuracy claimed on combined dataset was 94.97%.

B. Recognition of handwritten documents

1) *Online handwritten documents:*

C. C. Gohel et al. [13] in 2015 described the work on low-level stroke features for recognition of online handwritten Gujarati numerals and characters. The writers belonged to different educational background and different age group.

Researchers collected total 4500 samples, 1000 samples are of people with a lower educational background, 2500 samples are collected from students of schools and colleges, and last 1000 samples are collected from the highly qualified (i.e., Doctors, Professors, etc.). The recognition rate is 93%, 95%, 90% for Gujarati characters, Gujarati numerals, and for the combined dataset of Gujarati characters and numerals, respectively.

2) *Offline handwritten documents:*

In 2009, J. R. Prasad et al. [14] [15] worked on recognizing Gujarati handwritten characters. In this paper, researchers have generated a dataset for a handwritten character with 6 different classes with 10 different samples. Prasad et al. suggested a method known as pattern matching but the technique works efficiently for printed character recognition, not useful for handwritten character recognition. The performance evaluation for Gujarati handwritten characters shows an overall Recognition efficiency of 72%.

In 2010, Apurva A. Desai [16] developed a system to identify Gujarati handwritten digits, he has collected numerals 0-9 written in Gujarati language from 300 different people of various background and age group using a different pen, paper, even different writing style. Each data scanned at a resolution at 300 dpi using a flatbed scanner. His proposed method is used to recognize digits using ANN (Artificial Neural Network) and the system was able to achieve 82% accuracy for recognition of Gujarati digits.

In 2011, M. Maloo and K. Kale [17] described the work on affine invariant moments feature to recognize Gujarati numbers using SVM. They have collected datasheet from the random various 8 writers, belonging different educational background, age groups, and genders. Each datasheet scanned at a resolution at 300 dpi using HP 2400 Scanjet scanner. The overall maximum recognition rate 90.55% for SVM classifier.

In 2012, M. J. Baheti and K. Kale [18] generated a dataset of handwritten Gujarati numbers for identification of Gujarati Numbers. Data collected from different 16 people, belonging to age groups and various profession. They collected overall 1600 samples, where 10 different samples of each digit. They extracted features using the affine invariant moments and used classifiers namely Principal Component Analysis (PCA), Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Gaussian distribution function. The recognition rate is 84%, 92%, 90%, and 87% for PCA, SVM, KNN, Gaussian distribution function respectively.

In 2013, C. Patel and A. Desai [19] collected the data samples of all Gujarati alphabets from more than 200 writers. The writers belonged to various genders and age groups. They collected data from the writers who did not know about the Gujarati language. Each data scanned at resolution in-between 200-300 dpi using a flatbed scanner. In this work, researches proposed the hybrid features sets (i.e., Vertical line, No. of the object in one line, No. of the object in the upper/lower half, No. of the object in right/left half). They used tree classifiers and KNN classifiers in different stages and obtain low accuracy 63% by using hybrid features set.

In 2015, A. N. Vyas and M. M. Goswami [20] reported their work for handwritten numerals for Gujarati scripts. Researchers have collected the data samples from the 300 people of different age groups, educational background, and genders. They collected overall 3000 digits samples, where 900 digits written by thick marker pen, and other 2100 digits were written with a regular marker pen. In the data sample collected by them, all different kind of fonts, writing style size of digits is available. The authors have tested three different classifiers namely K-Nearest Neighbor, SVM (Support Vector Machine), and Backpropagation Neural Network. The recognition rate is 91%, 93%, and 92% for KNN, SVM, and Backpropagation NN respectively.

In 2015, M. M. Goswami and S. K. Mitra [21] proposed their work on handwritten numeral recognize using a low-level stroke. In this paper, researchers have generated a dataset for handwritten numeral with 14000 samples collected from 140 people with different age groups, different education background. They use K-nearest neighbor classifier and SVM (Support Vector Machine) classifier with Radial Basis Function (RBF) kernel. The recognition rate is 98.46% SVM classifier.

In 2015, S. J. Macwan and A. N. Vyas [22] reported work for Gujrati handwritten character. They have proposed a combination of three different methods namely Freeman chain code, Hu's invariant moment and center of mass which obtain accuracy rate of 87.22% for the proposed dataset.

In 2017, P. R. Paneri et al. [23] described the work on handwritten Gujarati word. They have used HOG (Histogram of Oriented Gradients) features and classifiers as SVM and KNN on the proposed dataset. The database consists of 2700 samples of handwritten Gujarati city name. They collected the 10 city names from 65 different age groups, different educational background with different pen, and papers. The sheets are scanned in 300 dpi at the flatbed scanner. The performance evolution for proposed dataset shown an average recognition rate of 85.87%.

IV. Analysis of the Literature Review

In this section, we reviewed various word recognition methods based on the literature and categorized them based on the types of dataset. Tables I, and II shows the analysis in brief.

Since there is no standard dataset available for the Gujarati language. There are around 39000 Gujarati books available online by DLI [24]. The government of Gujarat has also started many projects which transforms documents into a digital form such as e -Dhara project that transforms land record documents online in image format [25]. Thus, document image archival for printed as well as handwritten Gujarati documents are gradually increasing which demands an efficient document image retrieval system.

The first and traditional method is OCR systems, which is right now not available for handwritten, and historical documents as well as some degraded printed text documents. OCR systems presents several challenges in handwritten and printed Gujarati documents including, segmentation of word to character [26] and glyphs as well as the large and complex character set, join and cursive characters, modifiers, special symbols, overlapping characters, multiple writing styles and size of words, quality of documents, noisy and faded ink. Gujarati OCR gives high-level accuracy for character and numbers. Whereas word level accuracy for the Gujarati language remains low. Many works done is found in the classification step of the OCR system but very few works are found for the pre-processing step, segmentation and post-

processing steps of the OCR system. Hence, many researchers are motivated to explore text matching (also called word spotting) for document image retrieval. From 10 to 12 years word spotting plays an important role in image document retrieval tasks [28]. There are very fewer works reported for the literature review of retrieved images from the

TABLE I. HANDWRITTEN DOCUMENTS DATASET

Reference	Dataset					Features	Classifiers	Accuracy
	Categories	*Samples*	*Writers*	*Classes*	*Resolution*			
J. R. Prasad et al. [14] [15]	Character	10/ Character	-	6	-	-	Pattern Matching using Neural Network	72%
Apurva A. Desai [16]	Number	1600	300	10	300 dpi	Profile vector (Horizontal, Vertical, and two diagonal vectors)	Artificial Neural Network	82%
M. Maloo and K. Kale [17]	Number	800	8	10	300 dpi	affine invariant moments features	Support Vector Machine	91%
S. J. Macwan and A. N. Vyas [22]	Character	7800	-	-	300 dpi	Freeman Chain Code, Hu's Invariant Moment, Center of Mass	Support Vector Machine	87%
M. J. Baheti and K. Kale [18]	Number	1600	16	10	-	affine invariant moments features	Principle Component Analysis, Support Vector Machine, K-Nearest Neighbour, Gaussian distribution function	84%, 92%, 90%, 87%
A. N. Vyas and M. M. Goswami [20]	Number	3000	300	10	300 dpi	Modified Chain Code Method, Discrete Fourier Transform, Discrete Cosine Transform	Support Vector Machine, K-Nearest Neighbour, Backpropagation Neural Network	91%, 93%, 92%
M. M. Goswami and S. K. Mitra [21]	Number	14000	140	10	-	Low-level stroke features	Support Vector Machine with Radial Basis Function	98.46%
C. Patel and A. Desai [19]	Character	-	200	7 sets	200-300 dpi	Hybrid features	Tree classifiers, K-Nearest Neighbour	63%
P. R. Paneri et al. [23]	Word	2700	65	10 cities	300 dpi	Histogram of Oriented Gradients	Support Vector Machine, K-Nearest Neighbour,	86%
C. C. Gohel et al. [13]	Number, Character	3700, 1000	-	45	-	Low-level stroke features	K-Nearest Neighbour	≈93%

TABLE II. PRINTED DOCUMENTS DATASET

Reference	Dataset			Features	Classifiers	Accuracy
	Categories	*Samples*	*Classes*			
A. Antani and L. Agnihotri [27]	Character	40 sample/ class	10	First and higher order moment-based features	K-nearest neighbor and Minimum Euclidean distance	67%
J. Dholakia et al. [7]	Text	4,173	119	Wavelet features	K-nearest neighbor classifier and General regression neural network	96-97%
M. M. Goswami et al. [8]	Character	3000	32	Binary Vector	Self-Organizing Map, Projection with KNN Classifier	83%
M. Chaudhary et al. [9]	Character	Newspaper	6 sets	-	Multilayer Perceptron with Back Propagation Neural Network	96%
P. Solanki and M. Bhatt [10]	Document	748 images	-	Principal Component Analysis	Hopfield Neural Network	93%
E. Hassan et al. [11]	Character	16000	250	Histogram of Oriented Gradients, Feature Map, Modifiers Shape Descriptor features	Multiple Kernel Learning based Support Vector Machine Classifier	97-98%
M. M. Goswami and K. M. Suman [6]	Character	12000 symbols, 13000 symbols	42, 239	Low-level stroke features	K-nearest neighbor	98%, 95%
M. M. Goswami and S. K. Mitra [12]	Character	12000	42	High-level stroke features	K-nearest neighbor with shape similarity	95%

Gujarati printed documents [12] [29] using word matching.

Goswami et al. [12] presented the work for word matching on printed Gujarati documents. In proposed work, they extracted the high-level strokes from the printed Gujarati word documents using stroke extraction algorithm and retrieval of the matching word using shape similarity.

Another work found by Kathiriya et al. [29] in 2017, they used feature extraction techniques, namely Histogram of Oriented Gradient (HOG) and Shape Descriptors (SD). They performed word matching using DTW (Dynamic Time Wrapping) method on printed Gujarati documents and retrieval of the matching words from the document's image.

V. Conclusion

In this survey paper, a review of OCR work done on the Gujarati script is presented. Analysis of types of text recognition and the architecture of the OCR system is briefly discussed. There are various methods proposed to recognize the character, digits, and words so far for Gujarati script. Analysis of different methods concludes that character and digit level recognition have high accuracy, whereas word level recognition has low accuracy and work found on it is also less. The reason behind high accuracy is because there is single character and digit to recognize, whereas word level recognition is difficult due to a group of characters to recognize. From the study, we found that there is no standard dataset available for the Gujarati printed and handwritten documents. There is very less work reported for Gujarati language recognition, no literature work found using word matching for Gujarati handwritten documents. Therefore, the research efforts for document image retrieval for Gujarati script is justified.

References

[1] U. Pal and B. Chaudhuri, "Indian script character recognition: a survey," *Pattern Recognition,* no. 9, pp. 1887-1899, 01 09 2004.

[2] C. Cybersurf, "Central Intelligence Agency," 04 October 1994. [Online]. Available: https://www.cia.gov. [Accessed 23 October 2018].

[3] K. Harmandeep and K. Munish, "A comprehensive survey on word recognition for non-Indic and Indic scripts," *Pattern Analysis and Applications,* vol. 21, no. 4, p. 897–929, November 2018.

[4] Dalbir and S. K. Singh, "Review of Online & Offline Character Recognition," *International Journal Of Engineering And Computer Science,* vol. 4, no. 5, pp. 11729-11732, 5 May 2015.

[5] H. M. Kathiriya and M. M. Goswami, "Word Spotting Techniques for Indian Scripts: A survey," in *International Conference on Innovations in Power and Advanced Computing Technologies*, Vellore, India, 2017.

[6] M. M. Goswami and K. M. Suman, "Classification of printed Gujarati characters using low-level stroke features," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 15, no. 4, p. 25, 2016.

[7] J. Dholakia, A. Yajnik and A. Negi, "Wavelet Feature Based Confusion Character Sets for Gujarati Script," in *International Conference on Computational Intelligence and Multimedia Applications*, Sivakasi, Tamil Nadu, India, 2007.

[8] M. M. Goswami, H. B. Prajapati, and V. K. Dabhi, "Classification of Printed Gujarati Characters using SOM-based K-Nearest Neighbor Classifier," in *International Conference on Image Information Processing*, Shimla, India, 2011.

[9] M. Chaudhary, G. Shikkenawis, S. K. Mitra and M. Goswami, "Similar looking Gujarati printed character recognition using Locality Preserving Projection and Artificial Neural Networks," in *Third International Conference on Emerging Applications of Information Technology (EAIT)*, Kolkata, India, 2012.

[10] P. Solanki and M. Bhatt, "Printed Gujarati Script OCR using Hopfield Neural Network," *International Journal of Computer Applications,* vol. 69, pp. 33-37, 2013.

[11] E. Hassan, S. Chaudhury and M. Gopal, "Feature combination for binary pattern classification," *International Journal on Document Analysis and Recognition (IJDAR),* p. 375–392, December 2014.

[12] M. M. Goswami and S. K. Mitra, High-Level Shape Representation in Printed Gujarati Characters, vol. 1, SCITEPRESS, 2017, pp. 418-425.

[13] C. C. Gohel, M. M. Goswami and Y. K. Prajapati, "On-line Handwritten Gujarati Character Recognition Using Low-Level Stroke," *Third International Conference on Image Infonnation Processing,* December 2015.

[14] J. R. Prasad, U. V. Kulkarni, and R. S. Prasad, "Offline Handwritten Character Recognition of Gujrati Script using Pattern Matching," in *3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication*, Hong Kong, China, 2009.

[15] J. R. Prasad, U. V. Kulkarni, and R. S. Prasad, "Template matching algorithm for Gujrati Character Recognition," in *Second International Conference on Emerging Trends in Engineering & Technology*, Nagpur, India, 2009.

[16] A. A.Desai, "Gujarati handwritten numeral optical character reorganization through the neural network," vol. 43, no. 7, pp. 2582-2589, July 2010.

[17] M. Maloo and K. Kale, "SUPPORT VECTOR MACHINE BASED GUJARATI NUMERAL RECOGNITION," *International Journal on Computer Science and Engineering (IJCSE),* vol. 3, pp. 2595-2600, July 2011.

[18] M. J. Baheti and K. Kale, "Gujarati Numeral Recognition: Affine Invariant Moments Approach," *International Journal of electronics, Communication & Soft Computing Science & Engineering,* pp. 140-146, March 2012.

[19] C. Patel and A. Desai, "Gujarati handwritten character recognition using a hybrid method based on binary tree-classifier and k-nearest neighbor," *International Journal of Engineering Research & Technology (IJERT),* vol. 2, no. 6, p. 2337–2345, June 2013.

[20] A. N. Vyas and M. M. Goswami, "Classification of handwritten Gujarati numerals," in *International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, Kochi, India, 2015.

[21] M. M. Goswami and S. K. Mitra, "Offline handwritten Gujarati numeral recognition using low-level strokes," *International Journal of Applied Pattern Recognition,* vol. 2, no. 4, pp. 353-379, 2015.

[22] S. J. Macwan and A. N. Vyas, "Classification of Offline Gujarati Handwritten Characters," in *International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, Kochi, India, 2015.

[23] P. R. Paneri, R. Narang and M. M. Goswami, "Offline Handwritten Gujarati Word Recognition," in *Fourth International Conference on Image Information Processing (ICIIP)*, Shimla, India, 2017.

[24] D. Services, "National Digital Library of India," [Online]. Available: https://ndl.iitkgp.ac.in/. [Accessed 21 December 2018].

[25] R. Department and G. o. Gujarat, "E-Dhara," 03 June 1996. [Online]. Available: http://gil.gujarat.gov.in/edhara.html. [Accessed 21 December 2018].

[26] S. Chaudhari and R. Gulati, "Segmentation Problems in Handwritten Gujarati Text," in *International Journal of Engineering Research & Technology (IJERT)*, 2014.

[27] S. Antani and L. Agnihotri, "Gujarati Character Recognition," in *icdar*, 1999.

[28] W. G. A.-K. S. M. Rashad Ahmed, "A Survey on handwritten documents word spotting," *International Journal of Multimedia Information Retrieval,* vol. 6, no. 1, pp. 31-47, 2017.

[29] M. M. G. Himanshu M. Kathiriya, Performance Analysis of Word Spotting Techniques Using HOG and Shape Descriptor on Gujarati Script, vol. 671, Proceedings of the International Conference on Intelligent Systems and Signal Processing, 2018, pp. 163-168.